Hilary Zelko. Reasoning About Relevance. A Master's Paper for the M.S. in I.S degree. April, 2013. 51 pages. Advisor: Diane Kelly

This study focuses on how non-expert assessors judge relevance guided by mental models of relevance developed and applied during the assessment process. Components of relevance models are identified as well as challenges and changes associated with their construction and use. Study participants evaluated the relevance of news articles with respect to an assigned search topic. They commented on their reasoning in assessing each article, challenges they experienced in determining relevance and changes in their ability to assess relevance over the course of the evaluation session. Content analysis of these comments revealed that relevance models are derived from participants' understandings of the search topic, the documents they viewed and the relationships between them. Relevance manifestations (topical, situational, cognitive) and criteria (information scope, specificity and detail) guide the development and application of the relevance models, which may also be influenced by situational, cognitive and motivational factors.

Headings:

Information Retrieval

Relevance Assessment

Content Analysis

REASONING ABOUT RELEVANCE

by
Hilary Zelko

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2013

Approved by

_____

Diane Kelly

Table of Contents

Introduction and Literature Review

Ever since Cleverdon (1966) conducted the first evaluations of information retrieval systems in his Cranfield experiments, relevance has been a central concept used by IR researchers to understand the information search process as well as to design and evaluate IR systems. Relevance after all, is at the heart of information search, which is predicated on the identification and selection of information that is "relevant" to a searcher's information need or search task. But, what does it actually mean for information to be "relevant" to a need or task? How do people decide what is relevant and what are the most important factors that guide these decisions? Efforts to answer these questions have lead to a multitude of conceptualizations of relevance and no single all-encompassing definition or theory has been adopted, though IR researchers today accept that relevance is multidimensional, situational and dynamic (Borlund, 2003; Schamber, Eisenberg & Nilan, 1990; Saracevic 2007a; Saracevic 2007b).

1.1 Definitions and Conceptualizations of Relevance

Definitions and conceptualizations of relevance abound in the field of IR and extensive reviews have been written that analyze the different theories, types, aspects or manifestations of relevance (Schamber et al., 1990; Mizzaro, 1998; Cosijn and Ingwersen, 2000; Saracevic, 2007a; Hjorland, 2010; Huang and Soergel, 2013). In Saracevic's highly influential and widely cited review of the nature and manifestations of relevance (2007a), he draws on his earlier work as well as that of Cosijn and Ingerwersen (2000) and Borlund (2003) to summarize several different manifestations or types of

relevance including algorithmic, topical, cognitive, situational and affective. Algorithmic relevance is associated with the behavior of the search system that is designed to identify matches between queries and a corpus of documents. Topical relevance is concerned with the "aboutness" or subject matter relationship between a query and a corpus of documents. It can either be associated with the behavior of a system or the humans that interact with the system. Cognitive relevance refers to the relationship between a user's "cognitive state of knowledge" and the information within documents or other information objects. It is associated with criteria such as informativeness, novelty, and quality. Situational relevance is the relationship between the situation, task, problem-at-hand and the information objects retrieved. It pertains to things like usefulness for decision-making and uncertainty reduction. Finally, affective relevance is the relationship between the intents, goals, motivations of the user and the information retrieved. Saracevic noted that affective or motivational relevance may not be a separate manifestation, but rather underlie other types of relevance, particularly situational relevance.

Saracevic states, "relevance is a tangled affair involving interaction between and among a host of factors and variables" (2007a, p. 1926). He explains that there has been a divide in the research literature between system-based views of relevance and user-based views of relevance. The system view is primarily concerned with algorithmic or topical relevance, which is achieved by matching query terms or concepts to documents. The user view focuses on topical, cognitive, situational or affective relevance which may involve a variety of cognitive/psychological, or contextual factors that shape how human beings understand their search topics or information needs and the information objects

they encounter at a particular time and in a particular situation or context. In an effort to bridge the divide between these two perspectives, Saracevic proposed a stratified model that represents relevance "in terms of a set of interdependent, interacting layers..." (2007a, p. 1926) consisting of different "elements" or "processes" pertaining to humans and computers that shape the search process. Furthermore, different types of "relevances" are manifest in the different interdependencies or "relations" between these different strata (2007a). While the model has appeal for its integrative characteristics and has been embraced by many researchers, it has been difficult to operationalize and a challenge to specify exactly how these different types of relevances interact during the search process.

More recently, Huang and Soergel (2013) proposed a conceptual framework focused on topical relevance, which they argue "lies at the heart of" relevance (p. 18). Huang and Soergel emphasize the relational aspects of relevance discussed early on by Saracevic and incorporate situational and dynamic aspects identified by user-based researchers. Taking to heart Hjorland's (2010) criticism of the system/user dichotomy, they make an important distinction, not between systems and users, but rather between "relevance-as-is" and "relevance-as-determined" in order to "separate the conceptual definition of relevance from the measurement aspects of judging relevance or the operational aspects of computing relevance scores" (Huang and Soergel, 2013, p. 20). Relevance-as-is is a "relationship between an information object and a user's information need such that the information object has the potential of providing assistance in solving a problem, performing a task, producing a new document, learning about a given topic, satisfying curiosity, providing entertainment and so on" (p. 20). Relevance-as-is cannot be directly known and therefore can only be approximated by "relevance-as-determined"

which is “the result of the assessment or determination of relevance-as-is by a determining agent (person or computer system) based on representations of the information object and the information need made before, during or after use of the information” (p. 20). In Huang and Soergel’s (2013) model, information objects can include text or multimedia documents, images, speech, database tables, etc. and may be represented in a variety of ways such as title, author, abstract, passage, full-text etc. Information needs are complex and may include a topic or subject, user variables such as domain knowledge, search experience, cognitive style, problem/task variables such as purpose/intent/goal, task complexity, and situation/context variables such as situational constraints, and broader social/economic context. Determining agents may be people (subject matter experts, end users) or computer systems. The strength of this model is that it integrates a number of different conceptualizations of relevance that have been discussed in the literature and distinguishes between the definition of relevance and the process of determining relevance which depends on variable factors associated with representations of information needs and objects. In this study, the focus is on the relevance assessment process in an effort to better understand “relevance-as-determined.”

1.2 Relevance Assessment Criteria and Dynamics

Understanding the process of relevance assessment and the impact it has on the evaluation of IR systems has been an ongoing challenge in the field of IR. The experimental evaluation paradigm that began with Cranfield and evolved through Text REtrieval Conference (TREC) workshops has relied heavily on expert relevance

judgments and adopted a set of five core assumptions about relevance judgments (Saracevic, 2007b) that consisted of the following premises:

1. relevance judgments involve identifying a topical match between a query and information object
2. information objects are relevant or not relevant (binary)
3. relevance judgments can be made independently of one another
4. relevance judgments are stable and don't change significantly over time
5. relevance judgments are generally consistent and don't vary significantly across judges

Clearly, these assumptions are problematic and even early IR researchers (Cleverdon, 1970; Cuadra & Katter, 1967a; Cuadra & Katter, 1967b; Rees & Schultz, 1967) recognized that relevance judgments were variable and shaped by a wide variety of factors. Saracevic (2007b) documents a wide range of IR studies that challenge each of these assumptions. For example, he cites studies by Wang and Soergel (1998), Xu and Chen (2006) and Xu (2007) that found factors such as quality, novelty, and understandability to be important aspects of relevance judgments in addition to topicality. Other studies (Eisenberg, 1998; Janes, 1993; Spink, Greisdorf & Bateman, 1998) found evidence that people understand relevance along a continuum that includes degrees or levels of partial relevance. This nuance is not captured in binary relevance assessment and it has been suggested that graded or scaled measures of relevance should be used (Kekäläinen & Järvelin, 2002). A number of studies have found that relevance judgments are not independent and may vary depending on the ordering or size of documents presented (Eisenberg & Barry, 1988; Huang & Wang, 2004; Xu & Chen, 2006). In response to many of these challenges, IR researchers today are looking more closely at how much relevance assessments vary, what factors contribute to variation, and what

impact that variation has on relevance ratings and evaluation measures (Bailey, 2008; Yilmaz, 2012). A recent study of relevance assessment by Scholer, Kelly, Wu, Lee and Webber (2013) found that threshold priming (seeing varying degrees of relevant documents) impacted relevance assessments such that long sequences of irrelevant documents caused assessors to lower their thresholds and provide higher average relevance ratings than those who were exposed to highly relevant documents early in the assessment process. However, these effects diminished as people adjusted or "re-calibrated" their internal relevance models upon encountering documents with more diverse relevance levels. Scholer et al. (2013) also found a low level of self-agreement in ratings among individuals over time, which they suggested could either be a result of changes in peoples' internal relevance models or a result of other situational factors such as mental fatigue.

To better understand how the relevance assessment process actually works and how variations in relevance ratings come about, a number of researchers have examined relevance criteria that people draw on when assessing relevance. This area of research was inspired by the early work of Cuadra and Katter (1967) and Rees and Schultz (1967) who identified many factors that shaped relevance judgments of expert judges and was advanced by Schamber et al. (1990) who focused on how non-expert end-users evaluated relevance in the process of search. These studies lead to the identification of numerous variables that were associated with the information objects being evaluated (type, subject matter, level of difficulty), characteristics of the judges (experience, background, knowledge) and judgment conditions (time available, order of presentation, document size).

Schamber (1990, 1994), Park (1993), and Cool, Belkin and Kantor (1993) were among the first to explore relevance criteria among non-experts or "users" of IR systems as part of a larger effort to move outside of the traditional experimental paradigm and identify cognitive, dynamic and contextual factors that shape the interactions people have with systems during search. They identified a multitude of relevance criteria that were associated with user characteristics, the search topic/task, or the information objects/documents retrieved. Park (1993) identified 3 broad categories of factors including: 1) internal context or characteristics of users such as expertise in the problem area, previous research experience, education, 2) external context or aspects of the search such as goals or anticipated end product of search, stage of search, priority of information needs and 3) problem context or characteristics of the information problem such as intended use of the citation, and repetitiveness of information. Schamber (1991) grouped criteria based on aspects related to the information objects such as accuracy, specificity, and reliability. Cool et al. (1993) grouped criteria associated with both the user and the information objects. In an effort to identify the most important factors, Barry and Schamber (1995) compared criteria across their studies and identified areas of overlap including depth/scope/specificity, accuracy, clarity, and currency, which also corresponded with criteria, identified by Cool et al. (1993) and Park (1993). Recognizing the impracticality for further research of having so many different criteria and the lack of a consensus regarding the most important factors, Xu and Chen (2006) drew on Grices' theory of communication and earlier relevance work to conduct a factor analysis that identified a set of core relevance judgment criteria including, topicality, novelty, understandability, reliability and scope. Among these, topicality and novelty were found

to be the most significant factors. In a recent study using the TREC legal track, Chu (2011) found that topicality was the most important relevance criterion.

Relevance criteria have also been used to shed light on the dynamics of relevance assessment and "can be tied to changing cognitive states of users and changing situations involving users in the dynamic process of information retrieval" (Schamber et al., 1990). Taylor, Zhang and Amadio (2009) looked at how people may rely on different kinds of relevance criteria depending on their stage in the search process. For example, they found that people cited specificity as an important criterion early in the search process, while novelty become became more important later on. Taylor (2012) extended this work with a more detailed model of information search stages and found again, that different criteria were employed at different stages. However, for Taylor, changing relevance criteria are not important so much for what they say about the dynamics of relevance assessment per se, but rather that they reflect changes in peoples' cognitive states. He states, "As users retrieve documents, they make relevance judgments about documents reviewed based on various criteria. As the users' cognitive state changes, the criteria, which are important to their relevance judgments, may also change... Identifying associations between relevance criteria choices, relevance judgments, and search stage would provide insights into changes in the users' cognitive state" (2012, p. 136-137).

While relevance criteria are certainly important factors in relevance judgments and good indicators of changing cognitive states during search, the work focused on criteria so far does not really explain what these cognitive changes entail, how they shape a person's model of relevance and what contextual factors may influence that model

(Tang & Solomon, 1998; Zhang, 2008). How is a mental model of relevance formulated? How might that model change as people encounter and integrate new information?

1.3 Mental Models

Kenneth Craik first used the term mental model in 1943 to describe a "small scale" internal representation of reality used in human reasoning. The concept was later elaborated by Philip Johnson Laird who defined it as "...an iconic representation that is a structural, behavioral, or functional analog of a real-world or imaginary situation, event, object, or process" that is used to interpret and reason about the world (as cited in Nersessian, 2008). Mental models have been used in a wide variety of fields such as education, organizational behavior, and medicine to explain learning, reasoning and decision-making.

Mental models have been investigated in the human computer interaction (HCI) and information retrieval (IR) literatures primarily to understand how people learn and use information systems. This work has generally been directed toward improving the design of information systems as well as enhancing peoples' understanding and ability to use information systems. Borgman (1983) looked at how mental models could be used in training users about a system, which she later found enhanced people's search performance. In her study of web searching behavior, Slone (2002) found that mental models shaped peoples' search approaches, the web sites they visited and the sources they used. Westbook (2006) provided a theoretical overview of how mental models could be applied in information studies research more generally and presented results from an exploratory study that looked at patterns and components of mental models for information seeking among graduate students in a reference class. However, she adopted

a definition that focused on how people represented and modeled information systems and how those models impacted the information seeking process more generally. In a user study focused on how mental models were constructed during web-based search, Zhang (2008) found that models were constructed based on both peoples' internal cognitive states and external factors such as the system and the search task. She also found that mental models were dynamic and early models shaped the development of later models.

Tang and Solomon's (1998) study of the dynamics of relevance judgment is one of the few that explicitly uses the concept of mental models to explore how relevance judgments change during the search process. Using a naturalistic, case study approach with one searcher they explore from a "cognitive and situational perspective how relevance judgments evolve during the information retrieval process (1998, p. 254). While they find evidence of a "dynamic model of relevance" (p. 254) that evolved as their participant developed topical knowledge, they neither specify the components of the model nor how the relevance criteria fit in. Like Taylor (2012), they see changing relevance judgments and changing relevance criteria as indicators of "cognitive restructuring" but it is unclear what those cognitive structures consist of and how they might shape and be shaped by the relevance assessment process.

What are the components of a relevance model, how is it formed and transformed during the assessment process? Drawing on Huang and Soergel (2013) we can imagine a mental model of relevance that might involve a topic or need component based on an assessor's understanding or internal representation of the information need or a search topic, an object component based on one's understanding of the document or information

object encountered and a relationship component in which a relevance relationship is identified between the topic and object In the process of determining relevance, the relationship between topic and object is evaluated to determine the extent or degree to which the information encountered is relevant. If there is a large portion of content in the document that is related to the topic or need, then a higher relevance rating is likely, whereas documents with few contents addressing the topic would likely receive marginal or not relevant scores (Borlund, 2003). Of course with each encounter of a new document, there may be iterative modification of the model's topic component, the object component, or the relationship component, leading to adjustments in the relevance model as one progresses through a search. As Belkin (1982) has shown many search processes begin with "anomalous states of knowledge" in which a person has an ambiguous understanding of what they might be looking for and what kind of information may be relevant. As a result, they are likely to have a partially formed relevance model or possibly no model at all. As they develop topical knowledge, the relevance model may become clearer or more elaborate which could enhance a person's ability to identify a relationship (or lack thereof) between new information encountered and the topic of interest.

Because the relevance model may be adjusted as people encounter new information objects, it is likely that situational factors such as the type or sequencing of information presented may impact the contents, scope or application of the relevance model at any given time. For example, if documents that are highly relevant and rich in information about the topic are presented early in the process, a user may construct a more elaborate model earlier in the search process, which may cause them to reject or

marginalize documents with redundant information later on. This constitutes a "learning effect" (Xu & Wang, 2008; Harter, 1992) that may lead a person to use "more stringent topicality and novelty standards in judging documents in later stages as the user looks for more specific, more pertinent documents" (Xu & Wang, 2008, p. 1267). The learning effect may impact the degree of relevance that is determined at a given point in time as well as over time. On the other hand, as people read documents and learn about a topic, their information needs are satisfied and they could become less motivated to continue searching or reading about a topic. At the same time they are using up cognitive capacity, which could lead them to be less thorough in their assessment toward the end of the search process (Xu & Wang, 2008). This can result in what Xu and Wang (2008) refer to as a "cursoriness effect" in which documents encountered at the end of a search session might be evaluated on less stringent criteria because of fatigue, lower motivation and drained cognitive capacity that may set in during the course of a session. Therefore, the ordering of information during search may significantly shape how quickly and elaborately the relevance model is constructed and how it is applied at different points during an evaluation session.

Other situational factors such as peoples' existing knowledge about or interest in the search topic can affect how they judge relevance (Ruthven et al., 2007). Peoples' confidence in their relevance assessments may also be associated with the numbers of relevant documents identified and the level of relevance (Ruthven et al., 2007). The idea that situational factors may impact relevance judgments is not new. Drawing on Park (1992; 1993), Harter (1996) explains that because relevance judgments are shaped by the current state of a person's conceptualization of the information problem, they will vary as

that conceptualization is shaped by the citations encountered during search: “Thus, relevance of an individual citation [document, information object] is time-, order-, and situation-dependent” (1996, p. 39).

In this study, the focus is on how people develop relevance models that inform their decisions about relevance, the challenges they experience in developing or applying the models and the changes that occur in their models over time. In essence, that goal is to gain insight into “relevance-as-determined” (Huang and Soergel, 2013) by people during the process of assessing relevance by examining the reasons they use to explain their judgments as they evaluate a set of documents. These insights may help us to better explain the dynamics and variability of relevance assessments and guide how systems can be more effectively designed and evaluated.

Methods

The study was designed as a between subjects laboratory experiment with three conditions which varied in the number of non-relevant, marginally relevant and highly relevant documents that were presented. Initially one of the goals was to examine whether peoples' relevance models evolved differently according to condition, but due to time constraints and the large number of qualitative responses, the data were analyzed as a set rather than by condition.

All subjects were provided with one of three different search topics and then asked to evaluate a set of 48 newspaper articles with respect to the topic description. The three search topics, 385 (hybrid auto engines), 396 (sick building syndrome) and 415 (drug trafficking) were selected from a subset of the Trec-7 and Trec-8 collections developed by Sormunen (2002). The topic statements are included in Appendix A. Topics were chosen based on several factors. First, articles on the topic had to represent a sufficient mix of relevance levels. Second, topics that were unlikely to be familiar to study participants were chosen to minimize variation in existing knowledge participants might have about the topic. Finally, topics were chosen for their potential interest to study participants. Articles were presented one at a time and participants were asked to rate them using a 4-point categorical scale: not relevant, marginally relevant, relevant and highly relevant. Definitions for each relevance category were provided by the system.

The experimental conditions were embedded in the list of articles presented to study participants. Variations in the treatments occurred within the first twenty documents shown to participants. Ten of the documents were non-relevant documents presented in the same position (3rd, 5th, 8th) across all treatments. The other ten

documents were anchored in the same position, but varied across treatments. In the first treatment, 10 non-relevant documents appeared in these positions. In the second treatment, 10 marginally relevant documents appeared in these positions and in the third treatment, 10 relevant or highly relevant documents appeared in these positions. Participants were told that the documents were not in rank order according to relevance, but rather represented a random subset of documents retrieved by the search system. The 21st-48th articles shown to subjects were exactly the same, regardless of treatment and consisted of a mixture of documents with each of the 4 relevant rating types (non-relevant, marginally relevant, relevant and highly relevant) and three duplicate documents in positions 46, 47 and 48. The search topic and the documents came from a previously developed test collection, where several 'oracle judges' determined the relevance of the documents for each topic using the 4-point scale described above.

In addition to providing a relevance score, participants were asked to provide comments explaining their score and describe any changes they noticed about how they assessed the relevance of documents. These comments were analyzed using content analysis techniques in order to identify patterns and themes in users' relevance models during the evaluation.

A pre-test questionnaire was used to gather data for several contextual variables including familiarity with the topic, number of previous searches on the topic, interest in the topic and relevance of the topic to one's life. An exit questionnaire gathered basic demographic information (gender, age, level of schooling, major) and asked participants to comment on aspects of the relevance assessment process that were challenging as well as any differences they noticed in their assessments from the beginning to the end of the

evaluation. These comments were analyzed using qualitative content analysis techniques in order to better understand any factors that shaped the formulation and application of users' mental models of relevance.

2.1 Study Participants

Thirty-six study participants were recruited from the University of North Carolina Chapel Hill student body using a mass email to the student listserv. Participants were given a choice to enroll in one of six 1.5 hour evaluation sessions conducted over a period of one week. All participants were compensated $15.00 for their participation.

2.2 Data Collection

Upon arrival to the session, the researcher briefed participants on the goals of the study and ensured them that all data collected would remain confidential. Participants were given an informed consent form that acknowledged: that participation was voluntary, that participation might cease at any time and that the privacy of identification would be safeguarded. Each participant was given the URL for the online assessment system and a unique study ID and log-in. Participants were randomly assigned to treatments using a random number generator. The researcher addressed participant questions before the evaluation began.

Each participant completed a brief online training session built into the relevance assessment system to learn about the evaluation task and how to use the system. They then completed a pre-test questionnaire (see Appendix B) designed to gather information about previous searches they had conducted about the topic, knowledge and interest in the topic and the relevance of the topic to their lives.

As mentioned above, all participants were presented with 48 documents. They were asked to rate the relevance of each document on a 4-point categorical scale as well as select any portions of the documents that were relevant to the topic using a copy and paste function built into the system. In order to better understand how participants arrived at their relevance rating, all participants were given the following instruction:

Please enter comments about why you chose this particular relevance level for the document. You might like to include reflections about your decision-making process, or if you have any difficulties in coming to a decision. If you selected text segments in Step 2, you might like to include a short explanation of how they helped your relevance decision.

After assessing all documents, participants completed an exit questionnaire (see Appendix C) where they were asked to rate their confidence in their relevance judgments, explain any difficulties or challenges in determining the relevance of documents, as well as any differences in their assessment process between the beginning and the end of the session. Finally, participants provided basic demographic data including age, gender, student type (graduate, undergraduate, other), and major and whether or not they were native English speakers.

2.3 Participant Demographics

All participants except for two were native English speakers. The sample was composed of twenty-three females (64%) and thirteen males (36%). Twenty-eight (78%) participants were undergraduate students and eight (22%) were graduate students. Participant age ranged from 19 to 49 years old, though most participants (78%) were between the ages of 19 and 21. A wide variety of majors were represented across the following disciplines Science (19%), Social Science (38%), Humanities (10%), and

Professional School (33%). Some participants listed two majors, so they were counted once for each major, which resulted in a total of 42 responses.

2.4 Data Analysis

Qualitative content analysis is a research method used for interpreting text data through the "systematic classification process of coding and identifying themes or patterns" (Hsieh, 2005). It involves an inductive reasoning approach that allows categories, patterns or themes to "emerge from the data through the researcher's careful examination and constant comparison" (Wildemuth, 2009, p. 309). A major advantage of this approach is that it does not impose preconceived categories on the data and information generated from the analysis is based on participants' perspectives and grounded in the data (Hsieh, 2005).

The units of analysis in this study included participant comments explaining the relevance rating for each document as well as open-ended responses to two questions from the exit questionnaire that asked about challenges experienced in determining the relevance level and any differences experienced in assessing relevance from the beginning of the evaluation session to the end.

All comments were iteratively reviewed and coded for emerging themes and patterns. Multiple reviews of the data were conducted to identify codes and categories that emerged from the raw data and were not pre-determined. Categories were then examined to determine to what extent they exemplified or contradicted existing concepts and theories from the relevance literature.

Results

3.1 Prior Knowledge, Searches and Interest in the Topic

This section includes results from the pre-test questionnaire that asked about participants' prior knowledge, searches and interest in the topic. As shown in Figure 1, the vast majority of participants (92%) had little or no prior knowledge about the topic they were assigned.

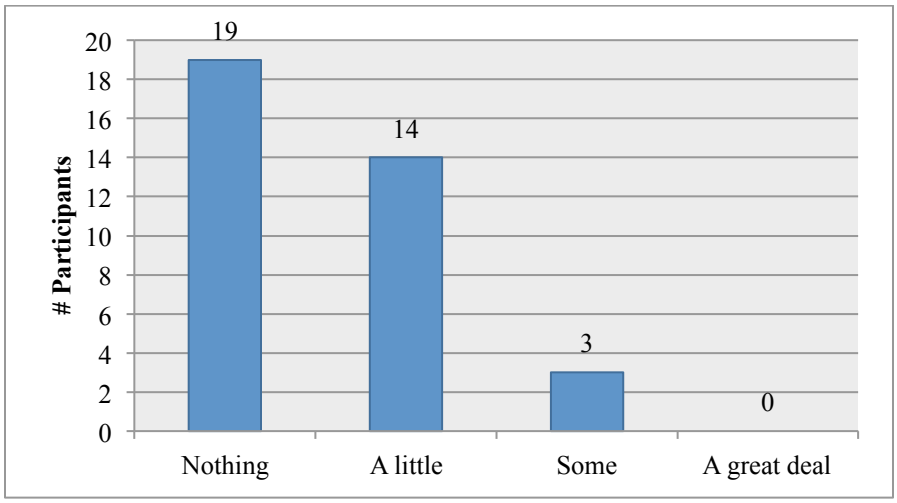

Figure 1: Knowledge about the topic

In addition to not knowing about the assigned topics, most participants (91%) had never searched for information about the topic as shown in Figure 2.

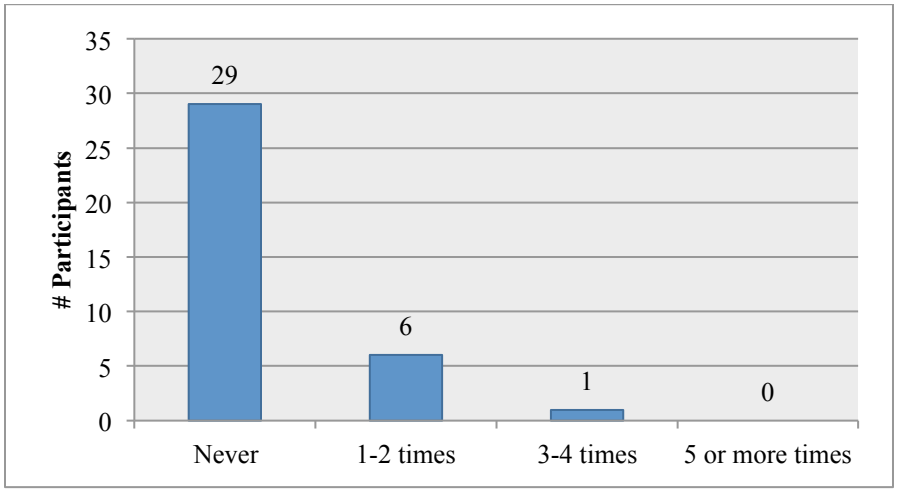

Figure 2: Past Searches on the topic

Despite having no knowledge and never having searched on the topics, a large majority of participants expressed some level of interest in the topic as shown in Figure 3, though most felt the topic was only slightly or not relevant to their lives as shown in Figure 4.

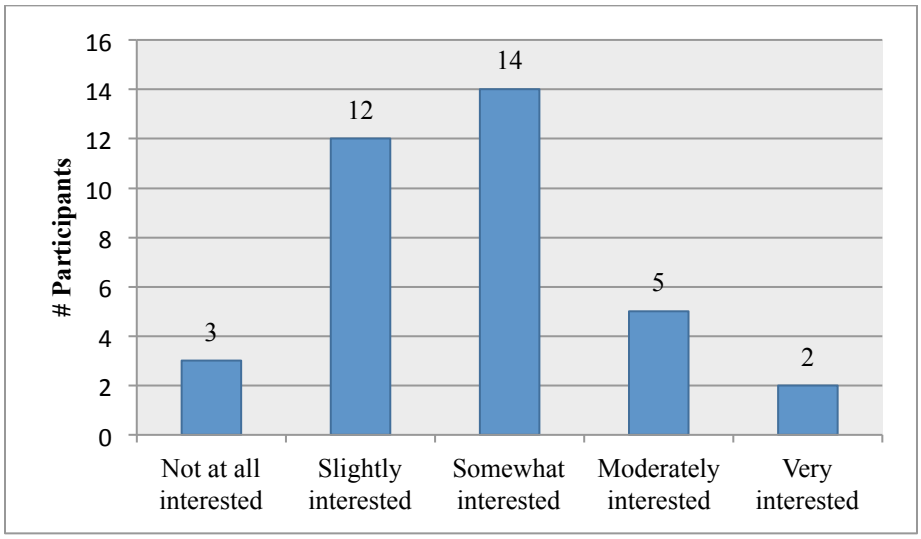

Figure 3: Interest in Topic

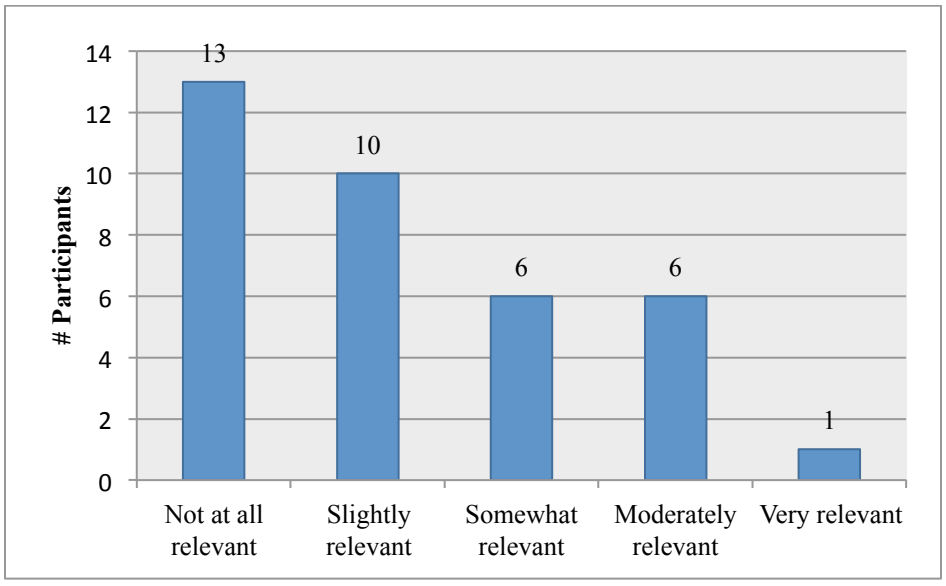

Figure 4: Relevance to ones' life

3.2 Confidence in Relevance Assessments

As part of the exit questionnaire administered at the conclusion of the evaluation session, participants were asked to rate how confident they were in the relevance judgments they made during the session. As shown in Figure 5, the majority of participants (58%) were moderately confident in their judgments while one third of participants were somewhat confident. Only one person felt highly confident in their judgments, which aligns with the qualitative findings that indicate the process of assessing relevance is far from straightforward and often challenging.

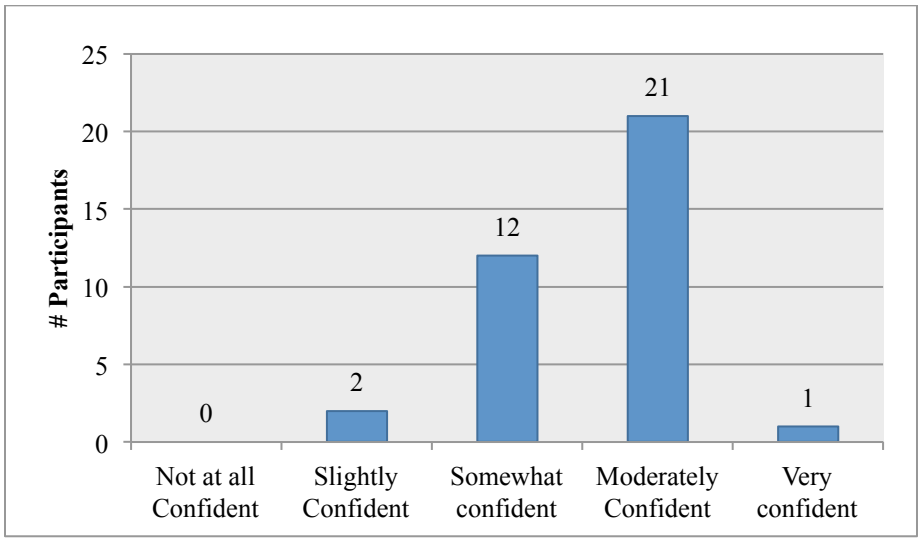

Figure 5: Confidence in Relevance Assessments

3.3 User Relevance Ratings

Participants rated each document using a categorical scale of Not Relevant, Marginally Relevant, Relevant and Highly Relevant. The distribution of each rating across all participants and all documents is shown in Figure 6. Overall, 45% of documents were rated as not relevant while 55% were rated as relevant to some degree.

Among those rated as relevant, documents were most frequently rated as “marginally relevant” (22%) and least frequently rated as “highly relevant” (14%).

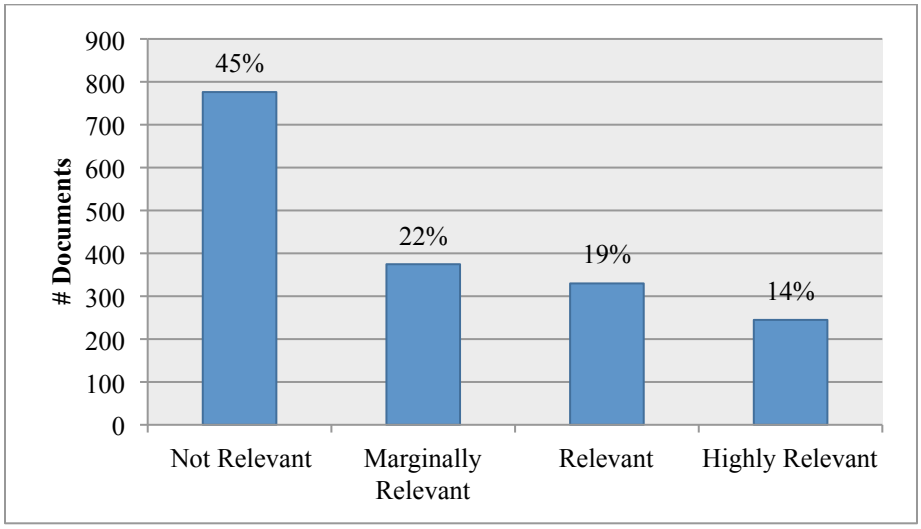

Figure 6: User Relevance Ratings

3.4 Reasoning about Relevance

In addition to rating the level of relevance of each document with respect to the specified search topic, participants explained the reasoning behind their rating. Each participant evaluated 48 documents for a total of 1729 documents in the study. While participants were asked to provide a comment for each document, the system allowed them to proceed to the next document without leaving a comment. There were a total of 1447 comments provided by all study participants. Overall, the comment response rates for each participant were high, suggesting that most participants took the task seriously. Twenty-four participants (67%) provided comments for at least 95% of the documents they evaluated, eight participants (22%) provided comments for 50-94% of documents

and four participants (11%) provided comments for less than 50% of the documents they evaluated.

Content analysis of these open-ended comments revealed several patterns in the way people model relevance that were shaped by their understanding of the topic and its various facets as well as the information presented in the documents and its relationship to those facets. While the study design emphasized topical relevance it was clear that participants drew on other kinds of relevance such as situational, cognitive and affective in making their assessments. Also, a few key relevance criteria including level of detail or specificity, information scope and information type were frequently cited in participants' reasoning about their relevance ratings.

It was evident from participants' explanations that their understanding of the topic statement was a key component of their relevance models. This topic component consisted of various aspects or facets that included a "main topic" (i.e. hybrid engines, sick building syndrome, drug trafficking) and various "subtopics" or "subthemes" (i.e. health effects of sick building syndrome, costs to consumers of hybrid engines), which were inferred from the specific keywords and concepts expressed in the topic statement. Documents were evaluated based on the extent to which they were "about" or "addressed/discussed" either the main topic and some or all of the different "subtopics". Aside from stopwords such as "and", "the" and "to" the term "about" was the most frequently used term in the explanations. Participants frequently specified which aspects of the topic were addressed and which were not. For example, one participant wrote, "It addresses most of the sub-topics with good detail but doesn't cover health benefits or trade-offs" to explain a partially relevant document. Another person stated, "…does

contain some relevant information about the topic of sick-building syndrome; however, the information only described the disease and how it occurs without giving relevant information about how it is affecting the employees." These kinds of explanations were coded in a category of partial information scope and were often associated with lower relevance ratings. Participants distinguished these partial scope documents from those that discussed most or all of the aspects or subtopics mentioned in the topic statement. For example, one participant noted that a document "touches on alternative fuel, cost to the consumer, comfort, horsepower, everything! Very relevant to the topic," while another person stated "this is a very relevant article because it has details about different car models, power trains, economical impacts and the outlook of alternative fuel vehicles." These kinds of explanations were coded in a category of full information scope and were associated with higher relevance ratings.

Another important pattern in the reasons associated with partial relevance assessments was reference to both topical and situational relevance. This was reflected in comments that indicated an article addressed some aspect of the topic, usually on a broad level, but in a different context than that specified in the topic statement, which made it less relevant. For example, one participant noted that an article "talks about hydrogen-fueled engines but within the context of aircraft engines, not automobile engines" while another article "talks about gasohol but not within the context of hybrid auto engines." Another participant stated, "Even though the document addresses drug trafficking, it does not mention the Golden Triangle which is the main point to look at." So even though an article may be "on topic" either in general or with respect to some subtopic, it must align with the context that was specified in the topic statement.

Overall, comments about the scope of information addressed in a document and its relationship to the topic and its various subcomponents were among the most frequent explanations for relevance assessments. Documents that addressed only some aspects of the topic were rated less relevant than those that addressed most or all of the aspects/subtopics/subthemes that were mentioned in the topic statement.

In addition to the information scope or context, the level of detail, specificity or "informativeness" of the information in the documents emerged as important criteria in assessing relevance. Participants frequently distinguished between documents that provided "good" or "adequate" detail, "thoroughly" or "exhaustively" "discussed", "made direct/clear reference to" the main topic or subtopics versus those that only "marginally touched on", "briefly" or "vaguely" "mentioned/referenced" the topic or some aspect of the topic. Other comments relating to this theme focused on how an article was "too broad" or "vague", provided only a "summary that doesn't give great detail", "doesn't offer anything substantial" or "talked about the broader issue, but gives no specifics." Several comments noted that articles "mentioned" or "talked about" a topic but "not in an informative way." One participant noted that an article, while on topic, would not be useful for a paper, "It is about drug trafficking, but I probably wouldn't use this in a paper I was writing on the Golden Triangle." Speaking of a highly relevant article one participant stated, "The entire article was relevant to the topic statement. It gave specific facts and direct quotes… I feel like I learned something through this article" while another stated "this article is full of useful information…" Overall, documents that provided a lot of detail, specifics, or were considered

"informative" or "useful" were associated with higher relevance ratings than those that did not.

The final major theme to emerge from the explanations involved the kind or type of information that was presented in the articles, which also impacted the relevance rating. For example, articles that discussed "pros and cons", "costs and benefits", "causes and effects" were often deemed relevant or highly relevant. Additionally, articles that "cited research", provided "sufficient examples" or provided "facts" or "data" were considered more relevant than those that were "opinion-heavy", expressed only "one persons perspective" or did not provide "in-depth analysis." Speaking of a highly relevant article, one participant stated "… Specific numbers and dates are given frequently through the article…" Interestingly, a couple participants noted that some articles were relevant because they provided high level or overview information even though they may not have had detail or lacked description. For example, one participant remarked that an article provided "an expansive review of electric car development in Europe" while another person stated an article was "highly relevant because it give an overview of the syndrome, although it is not very descriptive." On the other hand, documents that did not have enough "facts" or "evidence" were deemed less relevant as illustrated in the following comments: "it's on topic, but doesn't really give evidence", "offers a few facts, but not extensive", "…spoke about the search topic but offered little to no data regarding research about sick building syndrome." Some participants noted articles that only addressed "the legal side" or just talked about "regulations" were not considered very relevant. For example, one person stated an article "talks more about the legal ramification than the actual disease", while another stated, "contains information about

the legal side and does not include information about lung cancer a side effect of disease."

Overall, it was apparent from the explanations that topicality or "aboutness" was central to participants' relevance models, but topicality was multi-faceted which could explain multidimensionality in the relevance models. A key difficulty in assessing relevance was not knowing the exact boundaries of those facets which could also lead assessors to consider other kinds of relevance. Because assessors were given no information about situational relevance it was hard for them to know what kinds or types of information should be considered relevant, which was especially problematic in the face of vague, technical or specialized information. On the other hand, information that was highly "specific" or "detailed" helped to clarify those boundaries and made assessing relevance easier.

3.5 Challenges in Assessing Relevance Levels

The exit questionnaire included an open-ended question asking what challenges participants encountered when deciding the level of relevance associated with documents. A breakdown of responses is shown in Table 1. Since this was an open-ended question, some participants provided multiple challenges. Each of these responses was counted in the category in which it applied for a total of 44 challenges.

Challenges in Determining Relevance Levels	Frequency
Document was too long/short	9
Information scope or partial content- info too broad or only covers some aspects specified in the topic statement	9
Topic Ambiguity	7
Information Use or Type of information needed	6
Understandability - document was too technical or confusing	4
Redundancy of information	3
Lack of existing knowledge on topic	3
Stopped caring	1
No definition of relevance	1
Many factors in decision	1

Table 1: Challenges in Determining Relevance Levels

These challenges reflected difficulties that were associated not only with topical relevance, but also situational, cognitive and motivational relevance. The most common challenges expressed by participants were the length of the documents and the scope of information contained within them, which made it hard to assess topical relevance. Several participants mentioned that some of the documents were too long and they only skimmed long documents or found it difficult to "tease apart" information in a long document. On the other hand, short documents could also be difficult to assess if they did not contain enough information or the information presented was too "vague".

As with the explanations, information scope emerged as an important challenge in assessing relevance. Participants mentioned that it was especially difficult to determine the appropriate level of relevance when a document addressed only parts of the topic. For example, one participant stated, "it was hard to determine how relevant the articles were

when they were very close to being relevant. Some articles were completely about asbestos or pollution, but then never mentioned the health risks or how people working in exposed buildings can suffer as a result" and another said, "it was hard to determine whether I should consider the article relevant if it was just talking about the contaminants or what was wrong with a building but it didn't refer to any sicknesses caused by it." One person felt "compelled" to mark any document discussing a key figure mentioned in the topic statement as relevant even if the document did not address other aspects of the topic as illustrated in following statement, "The line between relevant and highly relevant was the most difficult to determine. Many documents discussed Khun Sa, who due to his position and reputation is intrinsically relevant to the topic statement, so I felt compelled to make almost any document involving him at least relevant, even if it did not go into detail on drug trafficking."

The next most commonly identified challenges were associated with evaluating both topical and situational relevance. Ambiguity about the terms and boundaries of the topic as well as uncertainty about the type of information that was needed made it hard to know how relevant a document was. Confusion or ambiguity about the keywords and concepts expressed in the topic statement was a challenge that made it difficult to determine the relationship between information presented and the topic. For example, one person mentioned they did not know exactly "what constituted 'data' about the diseases and illnesses", while another said, "the definition of hybrid was unclear. I wasn't certain if I should include electric cars, as they are not hybrids." One person noted that because the topic was "broad" it was "hard to distinguish if legal or examples were relevant" and another remarked he/she had to "frequently refer to the topic statement."

Speaking more generally, one participant stated that it was difficult not having a “clear definition about how each of these terms should be interpreted.”

Not having a clear sense of the topic terms and boundaries is related to another common challenge associated with situational relevance - not knowing how the information would be used. For example one participant stated, “the term relevance was used without reference to a specified goal e.g. medical treatment, investigative reporting, etc.” while another said, “It was difficult to determine whether slight irrelevant information that talked about drug trade elsewhere, or relations between the countries outside of drugs, was relevant. I didn't know exactly what the information was going to be used for, so that made it a little difficult.” The following statement expressed uncertainty over whether technical or “news” information was more important, “I wasn’t sure exactly what I was looking for. The prompt did not specify whether I should be looking for technical information on engines, or more "newsy" reports.”

Other challenges such as not having any background knowledge on the topic or not understanding the information presented reflected problems with cognitive relevance while difficulties assessing redundant information or losing interest and getting tired reflected motivational relevance. Several participants stated that not having any existing knowledge made it difficult to assess relevance and one person remarked that because of the lack of knowledge at the beginning it felt like “almost anything could be relevant.” Four people mentioned that some articles were too “technical” or “confusing” which made them hard to understand. Others noted that it was difficult to assess redundant information either because they did not want to “read through repetitive information” or because once “the information began to be repeated, it was hard to sort out what would

be relevant to someone just starting a search." Another participant explained that because there were many non-relevant articles at the beginning, it was hard to assess a "somewhat relevant" article when it appeared because it was unknown "how much more information was out there." One participant "stopped caring" while another said he/she would "skim" long and seemingly irrelevant articles to avoid getting too "tired" and not "understanding" things "later on." One person noted that there were "many different factors" in deciding between marginally relevant and relevant which could not be communicated.

3.6 Changes in Assessment Abilities over Time

The exit questionnaire also included an open-ended question asking participants to indicate whether they noticed a difference in their ability to assess relevance from the beginning to the end of the session. Almost all participants indicated that there was some change in their ability to assess relevance over the course of the session. These changes largely fell into two groups, those associated with a positive impact involving cognitive factors that produced a "learning effect" in which it became "easier" to assess relevance and those associated with a negative impact and a "cursoriness effect" in which motivational factors such as mental strain or loss of interest diminished the ability to assess relevance. As shown in Table 2, the majority of participants (18) indicated there was a positive effect on the ability to assess relevance from the beginning to the end of the session while a substantial number of participants (11) indicated that there was a negative effect. A few participants indicated that there was both a positive and negative effect, while two participants indicated that there was no difference and two participants declined to comment.

Changes in Assessment Ability over Time	Frequency
Positive impact – "improved", "easier", "better" over time	18
Negative impact – "tired", "bored" over time	11
Both positive & negative – "knew more, but became tired or stopped caring"	3
No difference	2
No answer	2

Table 2: Relevance Assessment Dynamics

Comments associated with a positive impact largely emphasized that it was "easier" to "decide" or "determine" relevance as they progressed and "learned/knew more" about the topic and "became familiar with the subject matter." Some comments suggested that with time they became more stringent in applying the relevance model and it became easier to "narrow down important information", "sort out extraneous info", "rule things out" or more easily identify/recognize "key phrases, topics and keywords" or "key people and places." Others suggested that "knowing more" and "having a better understanding" of the topic helped them to "know what to look for", "open up more" or "understand the language better and figure out the terms." A few participants indicated that they became more "confident" and could "skim" documents looking for keywords rather than "exhaustively going through each." A couple participants simply mentioned they "improved" or "performed better" by the end. While most participants indicated that knowing more about the topic had a positive impact on their ability to assess relevance, one participant mentioned that it was harder to determine relevance as they went "…because I became more familiar with the material, and was able to infer more about the article than it explicitly stated."

On the other hand, a number of participants felt that their abilities to assess relevance had diminished from the beginning to the end of the session largely due to factors associated with cognitive strain or losing motivation. A number of participants stated that they became "tired", "distracted" or "lost interest" by the end, which made it more difficult to "analyze relevance", "tease apart information", keep "focus" and "read through" documents, particularly those with "repetitive" or "monotonous" information. One participant mentioned, "my criteria slacked and I was more willing to connect something as relevant when it may not have been considered relevant toward the beginning." Another participant stated that it became harder to determine relevance as he/she became more familiar with the topic and that he/she may have "inferred more about the article than it explicitly stated."

A few participants suggested that there was both a positive and negative effect over time. While they had a better sense of "what to look for" over time, they also became "tired" or "bored." One person stated, "I think my ability peaked about half an hour into the session as I learned to sort through extraneous information, and dropped as the information became repetitive and I got fatigued."

Overall, it was clear that participants felt their ability to assess relevance can improve over time due to learning which presumably enables them to clarify their relevance models and apply them more confidently. On the other hand, the ability to assess relevance may also degrade over time as cognitive strain or boredom set in causing people to make errors, become careless or take shortcuts (i.e. skim or read only parts of the document) towards the end of an assessment session.

Discussion & Limitations

The study results provide further evidence that relevance is indeed multidimensional, situational and dynamic even within a controlled experimental setting. While participants were instructed to assess documents independently from one another, it was clear in their reasoning process that they struggled to do so. Also, using assigned topic statements, the study was designed with an emphasis on topical relevance, yet assessors frequently referred to other manifestations of relevance (situational, cognitive, motivational) when explaining their ratings and discussing the challenges and dynamics of relevance assessment. These results lend weight to Saracevic's assertion that relevance must be understood in terms of tangled, interacting "relevances" (2007a) that shape how people make decisions about relevance and shift over time.

Huang and Soergel's (2013) conceptual model of relevance is evident in the study findings in that people formulated models of relevance based on their understandings of the search topic, the documents and the relationships among them. The topic component of the model was based on the specific keywords and concepts expressed in the topic statement. It was multifaceted and consisted of a "main" or "broad" topic with subtopics or subthemes, which can in part explain the multidimensional aspects of relevance. The object component was formulated as participants reviewed articles and then compared it to the topic to see if there was a relevance relationship. The relationship was not an "all or nothing" affair, but varied based on the extent or degree to which the documents addressed the topic facets or whether they provided the right kind of information. Even when there was no relevance relationship found, participants often summarized the content in the document presumably to demonstrate that it did not relate to the topic.

Reading the documents caused participants to learn about the topic, which helped them to update and further develop their relevance models. It is likely that exposure to more highly relevant documents facilitated the development of clearer, more coherent models. This could lead these participants to have a higher relevance threshold resulting in lower average relevance ratings as was found in the study of threshold priming by Scholer et al. (2013). On the other hand, exposure to less relevant documents is likely to leave participants with less elaborate and more uncertain models that could lead them to a lower relevance threshold, resulting in higher average relevance ratings. Further analysis of the relevance ratings in this study could shed light not only on how threshold priming might impact the development of relevance models but the assessments themselves, and would contribute to the literature on the dynamics of relevance assessment.

Comments about the challenges associated with assessing relevance indicated that many participants had difficulty understanding or characterizing the relevance relationship, which caused them to wonder exactly how the information would be used. This suggests there might be a link between topical and situational relevance in the sense that unclear or complex relationships between information needs/topics and documents might lead people to consider how the information will be used rather than focusing on exactly what it is about. The study design could have amplified this finding because unlike a "real-world", user-initiated search scenario, participants had no basis on which to assess situational relevance since the topics were assigned and included no information about how the information would be used or why the search was being conducted. This could have contributed to the development of "fuzzy" models with respect to different types of relevance – topical, situational and cognitive (due to lack of existing knowledge

or exposure to many irrelevant documents), which made it especially difficult to determine the level of relevance and could have resulted in errors. These finding have implications for both the design and evaluation of information retrieval systems. With respect to design, it could mean that a system containing highly specialized or complex types of documents and search topics, may be easier to use if it presents views of that information organized around information use or users rather than purely on topicality or subject matching. With respect to evaluation, it suggests that in the context of assigned topics, it would be useful to provide assessors information relating to topical, situational and other kinds of relevance in order to facilitate the formulation of richer and more coherent relevance models.

Another outcome of this study was further evidence that different relevance criteria and manifestations shape the formulation and application of the relevance model. Topicality was central and it was evident that the "aboutness" of documents with respect to the topic statement was one of the most important components in applying the model and determining the rating. But "aboutness" can be multidimensional or multi-faceted particularly if topic concepts or boundaries are unclear, which could spur participants to consider other kinds of relevance such as situational or cognitive relevance in making their assessment. Information scope, specificity and detail were key criteria in determining relevance, and highly relevant documents were almost always accompanied by comments that emphasized they had good "detail" or fully covered all aspects of the topic. It seems likely that these criteria may be so important because they have direct bearing on all components of the model. That is, documents that are specific and detailed

are easier to understand, are likely to enhance one's understanding of the topic and could make it easier to recognize the relevance relationship.

Situational, cognitive and motivational factors also impacted participants' abilities to assess relevance. A number of participants mentioned challenges associated with ambiguity over the definition of terms in the topic statement, the definition of relevance itself, and lack of existing topic knowledge, which presumably made it difficult to construct the topic component and identify relationships between document information and the topic. It could be that concerns about how the information would be used (situational relevance) were related to the inability to formulate a clear topical relevance model, which may have been exacerbated by a lack of exposure to relevant documents that provided topical knowledge. On the other hand, exposure to relevant or highly relevant documents was likely to produce a strong "learning effect" that clarified the relevance model and made it easier to "recognize", "identify", "narrow down" relevant information and "rule out" extraneous information later on. While it was not mentioned as frequently, a substantial number of participants experienced the cursoriness effect and indicated that mental fatigue and loss of interest by the end of the session diminished their ability to assess relevance. Further research that explores the onset of these effects, how they interact and how they are impacted by the kinds and ordering of documents would be valuable in advancing our understanding of the dynamics of relevance assessment.

One of the main limitations of the study is inherent in the qualitative method used to analyze results. There is no prescribed "right way" to conduct inductive content analyses and the results often depend on the skills, insights and analytic ability of the

researcher (Elo, 2008). Also, because comments were completely open-ended and entered into an online system, some comments were inherently ambiguous, vague or uninterpretable and the researcher had no opportunity to follow-up or clarify what may have been meant by study participants. Although a systematic approach was used to code responses, and multiple reviews were conducted to ensure consistency, the large number of relevance explanations (~1450 comments) presented the possibility that some comments were miscoded or overlooked in the analysis process. Furthermore, because of time and financial constraints the study permitted only one analyst. Therefore, the reliability of results cannot be guaranteed.

Another limitation of the study was the fairly small and homogeneous sample that consisted entirely of college students. The results of the study may not be generalizable to the student population as a whole, people in general or other populations that are frequently involved in evaluating relevance such as expert assessors or non-academic users. Also, these student assessors were asked to articulate their reasoning for each rating which may have cause them to behave differently than assessors who are asked only to provide a rating.

Finally, while the experimental, lab-based study design was important in order to control the conditions of assessment process, the assignment of topics and the requirement that participants read a large number of documents during a lengthy, single session are unlikely to match the shorter, user-initiated, multi-session searches that are more typical of the student population. However, the findings of the study could be useful for understanding relevance assessment in other contexts such as TREC collection

development, machine learning or legal research settings in which assessors have assigned topics and are asked evaluate numerous documents in a single session.

Conclusion

The purpose of this study was to explore how people reason about and model relevance during the relevance assessment process as well as to identify factors that may challenge or alter their ability to do so. More specifically, the study aimed to identify different components of and influences on the relevance model by examining how people explain their assessments, the challenges they encounter and changes in assessment abilities over time. Qualitative content analysis methods were used to evaluate over 1500 open-ended comments describing reasoning processes, challenges and changes in relevance assessments made by student assessors during lab-based evaluation sessions.

The study results show that while relevance models vary across individuals and change over time, they are shaped by several core relevance manifestations and criteria and have common components derived from information topics/needs, documents and their relationships. Challenges in assessing relevance were spread across these components. Some participants found it most difficult to understand the terms and concepts of the topic and where its boundaries lay, some had trouble sorting out information in the documents and others struggled with understanding the relationship.

In reasoning about their relevance assessments, participants cited factors associated with different relevance manifestations and key criteria such as scope, context, detail or specificity and type of information, which aligns with existing research on relevance. These criteria may work via the different components of the relevance model and impact how it is formulated, modified and applied as people learn and encounter new information. For example, encountering highly detailed or specific information may help to clarify the topic component, which may lead people to apply the relevance model more

stringently when they encounter broad or vague information later on. Most participants experienced a "learning effect" that made assessing relevance "easier" because they could better distinguish between "extraneous" and relevant information. However, this could be counterbalanced by a "cursoriness effect" driven by mental strain or loss of interest in the task.

Future research investigating interactions among learning and cursoriness effects during relevance assessment as well as factors that might contribute to their onset could advance our understanding of the relevance assessment process. One important factor could be the ordering or sequencing of documents (threshold priming) which not only might impact the onset of these effects but is also likely to shape the development and application of the relevance model. Finally, the fact that the study participants were asked to articulate their reasoning for each relevance assessment might have caused them to be more aware of and more fully develop their relevance models than they would have if asked only to provide a relevance rating. Therefore, it would be useful to compare assessments made under these different conditions to determine whether making the modeling process explicit has an effect on relevance assessments

Bibliography

Bailey, P., Craswell, N., Soboroff, I. Thomas, P., deVries, A.P., & Yilmaz, E. (2008). Relevance assessment: Are judges exchangeable and does it matter? *Proceedings of the 31st international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '08),* 667-764.

Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management,* 34(2/3*)*, 219-236.

Belkin, N. J., Oddy, R.N., & Brooks, H. M. (1982). ASK for information retrieval: Part I. Background and theory. Part II. Results of a design study. *Journal of Documentation*, 38, 61-71.

Borlund, P. (2003) "The concept of relevance in IR," *Journal of the American Society for Information Science*, vol. 54, 913–925.

Borgman, C.L. (1986) The user's mental model of an information retrieval system: an experiment on a prototype online catalog, *International Journal of Man-Machine Studies*, Volume 24, Issue 1, January, 47-64.

Cleverdon, C. W., & Keen, M. (1966). Factors determining the performance of indexing systems. Vol. 1: Design, Vol. 2: Results. Cranfield, Bedford: Aslib Cranfield Research Project.

Chu, H. (2011). Factors affecting relevance judgment: a report from TREC Legal track. *Journal of Documentation*, Vol. 67 Iss: 2, 264 – 278.

Cool, C., Belkin, N. J., & Kantor, P. B. (1993). Characteristics of texts affecting relevance judgments. In M.E, Williams (Ed.) *Proceedings of the 14th National Online Meeting* (pp. 77 84). Medford, N J: Learned Information.

Cosijn, E. & Ingwersen, P. (2000) Dimensions of relevance. *Information Processing & Managemen*t, Volume 36, Issue 4, 533-550.

Cuadra, C. A., & Katter, R. V. (1967). Experimental Studies of Relevance Judgments Final Report. Volume 1: Project Summary. Cleveland, OH: Case Western Reserve University, School of Library Science, Center for Documentation and Communication Research.

Eisenberg, M., & Berry, C. (1988). Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science,* 39(5), 293-300.

Elo S., & Kynga, S H. (2008) The qualitative content analysis process. *Journal of Advanced Nursing* 62(1), 107–115.

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13), 1120-1129.

Janes, J. W. (1993). On the distribution of relevance judgments. *Proceedings of the American Society for Information Science*, Columbus, Ohio, 104-114.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9), 602-615.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science,* 47(1), 37-49.

Hjorland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology,* 61(2), 217-237.

Hsieh, H., & Shannon, S. (2005) Three Approaches to Qualitative Content Analysis. *Qual Health Res November* 15: 1277-1288.

Huang, X., & Soergel, D. (2013). Relevance: An improved framework for explicating the notion. *Journal Of The American Society For Information Science & Technology*, *64*(1), 18-35.

Huang, M., & Wang, H. (2004). The influence of document presentation order and number of documents judged on users' judgments of relevance. *Journal of the American Society for Information Science and Technology, 55*(11), 970-979.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48*(9), 810-832.

Nersessian, N. J. (2008). *Creating scientific concepts*. Cambridge, Mass.: MIT Press.

Park, T. K. (1993). The nature of relevance in information retrieval: an empirical study. *The Library Quarterly*, 63, 318-351.

Park, T. K. (1994). Toward a theory of user-based relevance:A call for a new paradigm of inquiry. *Journal of the American Society for Information Science, 45*(3), 135-141.

Rees, A.M., & Schultz, D.G. (1967). A field experimental approach to the study of relevance assessments in relation to document searching: final report: volume 1. Cleveland, OH: Case Western Reserve University, School of Library Science, Center for Documentation and Communication Research.

Ruthven, I., Baillie, M. & Elsweiler, D. (2007). The relative effects of knowledge, interest and confidence in assessing relevance. *Journal of Documentation*, Vol. 63 No. 4, 482-504.

Saracevic, T. (2007a). Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, Vol. 58 No.13, 1915-33.

Saracevic, T. (2007b). Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, Vol. 58 No.13, 2126-44.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management,* 26(6), 755-776.

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29, 3–48.

Scholer, F., Kelly, D., Wu, W.C., Lee, H., & Webber, W. (2013, under review). "The Effect of Threshold Priming and Need for Cognition on Relevance Assessment." *ACM SIGIR 2013 Conference on Research and Development in Information Retrieval*.

Slone, D. J. (2002). The influence of mental models and goals on search patterns during web interaction. *Journal of the American Society for Information Science and Technology*, 53(13), 1152-1169.

Sormunen, E. (2002). Liberal relevance criteria of TREC – counting on negligible documents? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, 324-330, Tampere, Finland.

Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: examining different regions of relevance. *Information Processing and Management,* 34(5), 599- 624.

Tang, R., & Solomon, P. (1998). Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information Processing and Management*, 34(2/3), 237-256.

Taylor, A., Zhang, X., & Amadio, W. J. (2009). Examination of relevance criteria choices and the information search process. *Journal of Documentation*, 65(5), 719-744.

Taylor, A. (2012). User relevance criteria choices and the information search process. *Information Processing and Management*, 48(1), 136-153.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5), 697-716.

Wang, P., & Soergel, D. (1999). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science*, 49(2), 115–133.

Westbrook, L. (2006). Mental models: a theoretical overview and preliminary study. *Journal of Information Science* , 32 (6), 563–579.

Wildemuth, B. M. (2009). *Applications of social research methods to questions in information and library science*. Westport, Conn.: Libraries Unlimited.

Xu, Y. (2007). Relevance judgment in epistemic and hedonic information searches. *Journal of the American Society for Information Science and Technology*, 58(2), 178–189.

Xu, Y. C., & Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology, 57*(7), 961-973.

Xu, Y., & Wang, D. (2008). Order effect in relevance judgment. *Journal of the American Society for Information Science and Technology*, 59(8), 1264-1275.

Yilmaz, E. Kazai, G., Craswell, N., & Tahaghoghi, S.M.(2012). On judgments obtained from a commercial search engine. *Proceedings of the 35th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '12),* 115-116.

Zhang, Y. (2008). The influence of mental models on undergraduate students' searching behavior on the Web. *Information Processing & Management*, 44 (3), 1330–1334.

Appendix A: Topics

1. Hybrid Auto Engine (385)
   Identify documents that discuss the current status of hybrid automobile engines, (i.e. cars fueled by something other than gasoline only).

   A relevant document may include research on non-gasoline powered engines or prototypes that may be fueled by natural gas, methanol, alcohol; cost to the consumer; health benefits derived; and shortcomings in horsepower and passenger comfort.

2. Sick Building Syndrome (396)
   Identify documents that discuss sick building syndrome or building-related illnesses.

   A relevant document would contain any data that refers to the sick building or building-related illnesses, including illnesses caused by asbestos, air conditioning, pollution controls. Work-related illnesses not caused by the building, such as carpal tunnel syndrome, are not relevant.

3. Drugs Golden Triangle (415)
   What is known about drug trafficking in the "Golden Triangle", the area where Burma,Thailand and Laos meet?

   A relevant document will discuss drug trafficking in the Golden Triangle, including organizations that produce or distribute the drugs; international efforts to combat the traffic; or the quantities of drugs produced in the area.

Appendix B: Pre-Test Questionnaire

1. How many times have you searched for information about this topic in the past?
    - Never
    - 1-2 times
    - 3-4 times
    - 5 or more times

2. How much do you know about this topic?
    - Nothing
    - A little
    - Some
    - A great deal

3. How interested are you to learn more about this topic?
    - Not at all interested
    - Slightly interested
    - Somewhat interested
    - Moderately interested
    - Very interested

4. How relevant is this topic to your life?
    - Not at all relevant
    - Slightly relevant
    - Somewhat relevant
    - Moderately relevant
    - Very relevant

Appendix C: Exit Questionnaire

1. How confident are you in the relevance judgments you made?
    - Not at all confident
    - Slightly confident
    - Somewhat confident
    - Moderately confident
    - Very confident

2. What, if anything, was challenging about deciding which relevance levels (not relevant, marginally relevant, relevant and highly relevant) to associate with each document?

3. Did you notice any differences in your ability to determine relevance from the beginning of the end of the session? Please explain.

4. Sex
    - Male
    - Female

5. Age

6. Are you a:
    - Graduate student
    - Undergraduate student
    - Other

7. What is your major course of study?

8. Is English you native language?
    - Yes
    - No